Dogs perceive and spontaneously normalise formant-related speaker and vowel differences in human speech sounds

Article (Accepted Version)

Title: Dogs perceive and spontaneously normalise formant-related speaker and vowel differences in human speech sounds

Authors: Holly Root-Gutteridge*[a], Victoria F. Ratcliffe[b], Anna T. Korzeniowska[a], David Reby[a]

Affiliations:

[a] Mammal Vocal Communication & Cognition Research Group, School of Psychology, University of Sussex, Brighton BN1 9QH, UK

[b]Defence Science and Technology Laboratory, Salisbury, Wiltshire, UK

[c] Equipe de Neuro-Ethologie Sensorielle ENES / CRNL, University of Lyon / Saint-Etienne, CNRS UMR5292, INSERM UMR_S 1028, Saint-Etienne, France

*Corresponding author email: hollyrg@googlemail.com

Abstract

Domesticated animals have been shown to recognise basic phonemic information from human speech sounds and to recognise familiar speakers from their voices. However, whether animals can spontaneously identify words across unfamiliar speakers (speaker normalisation) or spontaneously discriminate between unfamiliar speakers across words remains to be investigated. Here, we assessed these abilities in domestic dogs using the habituation-dishabituation paradigm.

We found that while dogs habituated to the presentation of a series of different short words from the same unfamiliar speaker, they significantly dishabituated to the presentation of a novel word from a new speaker of the same gender. This suggests that dogs spontaneously categorised the initial speaker across different words. Conversely, dogs who habituated to the same short word produced by different speakers of the same gender significantly dishabituated to a novel word, suggesting that they had spontaneously categorised the word across different speakers. Our results indicate that the ability to spontaneously recognise both the same phonemes across different speakers, and cues to identity across speech utterances from unfamiliar speakers, is present in domestic dogs and thus not a uniquely human trait.

Keywords: speaker normalisation; vowel perception; speaker discrimination; speech perception

Background

Speech sounds vary among speakers due to differences in body size, age, gender, and other idiosyncratic attributes [1,2], and thus effective speech perception relies on a listeners' ability to recognise phonemes independent of such speaker variability, a perceptual mechanism known as speaker normalisation [3]. In human speech, vowels are represented by specific formant frequency patterns, but the absolute values of the formants vary across speakers due to size-, age- or other individual differences in vocal tract length [4,5]. Yet these speaker-related differences in formant values encode socially relevant indexical and identity cues across phonemes [6,7]. Thus, human listeners must normalise these two dimensions of speech variation to recognise words across different speakers and to identify individual speakers across different words, an ability that was once posited to be uniquely human [8]. Although some nonhuman animals can be trained to recognise phonemes across speakers and have also been shown to recognise familiar humans from their voices [review: 8], both the extent to which animals can spontaneously perform speaker normalisation to recognise words across unfamiliar speakers and their ability to spontaneously discriminate between unfamiliar speakers across speech sounds remain to be investigated.

Here, we use domestic dogs (*Canis familiaris*) to investigate these abilities in a nonhuman mammal that is regularly exposed to human speech utterances that function as interspecific signals. Indeed, dogs are known to recognise basic phonemic information, for example, when following commands (even in the absence of tonal cues [10,11]), and can recognise familiar human voices speaking known phrases [12,13]. However, in order to recognise words across speakers, dogs must attend to the relative positions of formants in human speech rather to than their absolute values by normalising variation in the acoustic signal that is related to speaker identity or gender [14]. Moreover, to discriminate between

56 unfamiliar speakers, dogs must also be able to attend to these same speaker cues across
57 different phonemes. As performing one task could preclude the other, we investigated
58 whether dogs would spontaneously normalise variation in human speech to recognise
59 words across speakers, and speakers across words, using the habituation-dishabituation
60 paradigm. This paradigm has been used widely in perceptual studies involving animal or
61 nonverbal participants [15–17], and has been used previously to explore dogs' ability to
62 discriminate conspecific barks produced by different individuals [18].
63 To investigate dogs' ability to spontaneously discriminate between unfamiliar
64 speakers, we tested whether dogs would habituate to a short series of different single
65 syllable words [i.e. *H*-vowel-*D*] that varied only in the vowel and were produced by the same
66 unfamiliar speaker, then dishabituate to the presentation of a new [*H*-vowel-*D*] word from a
67 different speaker, then re-habituate to a final novel [*H*-vowel-*D*] word from the original
68 speaker (Electronic supplementary materials: Figure 1A). We predicted that if the dogs
69 spontaneously categorised the identity of the initial speaker across words and recognised a
70 change in speaker, then they would show a longer response to the dishabituation stimulus
71 word than to the final habituation or re-habituation stimuli words.
72 Next, we investigated dogs' ability to spontaneously normalise voice differences
73 across speakers in order to discriminate between phonemes. We exposed them to four
74 examples of the same word produced by four unfamiliar, same-gender speakers, then
75 introduced a new speaker producing a new word (ESM: Figure 1B). We predicted that if the
76 dogs spontaneously categorised the word produced by the different speakers, then they
77 would show an increase in response duration to the dishabituation stimulus, demonstrating
78 that they recognised the change in word and had spontaneously normalised production
79 across speakers.

Methods and materials

Voices from 13 adult men and 14 adult women who were not familiar to the dogs were sampled with a randomised presentation of voices across conditions. We used four habituation, one dishabituation, and one re-habituation sound stimulus trials with 6 seconds of silence between each audio stimulus presentation [19]. Speaker identity and order of presentation of vowels were all pseudo-randomised across stimuli. For further details, see ESM Methods.

For trials in condition 1 (speaker discrimination), the discrimination of unfamiliar voices was tested with sequences using the voices of four unfamiliar speakers who produced monosyllabic words. Each stimulus word started with “h” and ended in “d” following [20], and included one of nine vowel-sounds: “had”, “head”, “heard”, “heed”, “hid”, “hod”, “hood”, “whod”, and “hud”. In condition 2 (speaker normalisation), the discrimination of the vowels [a], [i], and [o] was tested using “had”, “hid” and “whod”. These vowels were chosen and paired so as to be clearly distinct from one another and difficult for dogs to confuse. In both conditions, half of the stimulus sequences involved female voices and the other half involved male voices. While these short words may be familiar to dogs, they are not typically used in commands in the English language.

A total of 70 dogs participated in the between-subject design study. Each dog heard 6 sounds, with 24 dogs retained in each of the two conditions (see ESM for demographic details). Videos were assessed before coding and discarded if the dog either did not visibly respond to the stimulus by moving any part of their face or body including their eyes (n=4 dogs) or was distracted during trials by non-stimulus sounds or events (n=18). The stimuli

were presented from an Apple iBook Air through a Behringer Europort MPA40BT-PRO speaker that was set to conversational volume (approx. 65 dB) and placed on one side of the dog, counterbalanced across subjects. The dogs' reactions were filmed on a Sony FDR-AX100 camcorder positioned on a tripod. Duration was measured as the time between the initial onset of response (e.g. looking, ears moving into forward position, eyes looking in direction of the speaker, head turning, or moving towards the speaker), until the dog stopped visibly responding or the beginning of the next trial. All abovementioned responses were coded as "change in behaviour". Lack of response was coded as duration equals zero. All videos were coded blind in Sportscode Gamebreaker 11 (Sportstec, Warriewood, NSW, Australia) by HRG with 25% double-coded blind by ATK (see ESM for details).

Statistical tests were performed in SPSS v. 25 (SPSS Inc., Chicago, IL., USA). Linear mixed effect models (LMEs) fitted with restricted-maximum likelihood estimation were used to examine the effect of trial on listener response duration. Dog identity was included as a random effect and fixed effects included trial, dog sex, age in years, breed-group, recording location, and speaker-gender, with significance threshold calculated at $p < 0.007$ using Bonferroni to correct for multiple comparisons. The variables met LME assumptions and residuals were normal as indicated by Shapiro-Wilks tests.

Results

Duration of the dogs' responses in each trial was not significantly different across conditions ($F_{1,187.5} = 5.961$, $p = 0.016$, with corrected threshold of $p = 0.007$). For both conditions, only the habituation trial factor had a significant effect on response duration, while there were no other significant fixed effects ($p > 0.05$ for all other variables, see ESM for details).

a)

b)

Figure 1 Boxplots of duration of response to stimulus sounds for a) Condition 1: speaker discrimination (n = 24 dogs), and b) Condition 2: speaker normalisation (n= 24 dogs). P values < 0.05 marked by *, p < 0.01 marked by **, p < 0.001 marked by ***, and outliers are marked by circles. H = Habituation trial, DH = dishabituation trial, RH = Re-Habituation trial.

The LME results were similar for both conditions: habituation trial had a significant effect on response duration (condition 1, speaker discrimination: $F_{5,115} = 4.271$, p = 0.001, condition 2, speaker normalisation: $F_{5,115} = 5.421$, p < 0.001). Response duration decreased in both conditions from habituation trial 1 to trial 4 (condition 1: p = 0.047; Figure 1a; condition 2: p = 0.001, Figure 1b), showing that dogs habituated to the stimuli over time.

For both conditions, dogs' response durations increased significantly for the dishabituation trial compared to final habituation trial 4 (condition 1: p = 0.007, condition 2: p = 0.001) and the re-habituation trial (condition 1: p = 0.001, condition 2: p < 0.001), showing that they dishabituated to the change in stimulus and re-habituated to the repeated stimulus. Response duration in the re-habituation trial was not significantly different to the final habituation trial 4 (condition 1: p = 0.413, condition 2: p = 0.778) while the dishabituation trial response duration was not significantly different to habituation trial 1 (condition 1: p = 0.467; condition 2: p = 0.953). Thus, the duration of dogs' responses to the dishabituation trial was similar to that of their original response to the first stimulus.

Formatte

Formatte

Deleted: F

These results show that dogs habituated to the same speaker producing four different words dishabituated to a new speaker producing a new word (Figure 1a). This demonstrates that dogs can spontaneously categorise short words as belonging to the same unfamiliar speaker based on the presentation of a very limited set of four stimuli, and are thus able to detect a change in speaker identity when a new speaker produces a new word that was not used in the habituation sequence. Conversely, dogs habituated to the same word spoken by four different speakers of the same gender and then dishabituated to a new word spoken by a new speaker which differed only in its vowel, demonstrating that dogs detected a change in the vowel sound, which can only be achieved by categorising the vowels as similar in the habituation sequence, despite speaker differences in formant frequencies (Figure 1b).

Deleted: F

Deleted: F

Discussion

Our results provide the first demonstration that spontaneous speaker normalisation is not unique to humans, as we show that domestic dogs can spontaneously discriminate the same words across speakers. We also show that dogs are capable of spontaneously discriminating between unfamiliar speakers of the same gender across different words, suggesting that they have the ability to extract identity information from unfamiliar human voices on the basis of very little acoustic exposure. As interindividual differences in pitch were removed from vocal stimuli, dogs could only discriminate the speakers based on filter-related cues common to the different vowels, and/or on subtle idiosyncratic information encoded in the surrounding consonants.

Previous work on speaker normalisation in non-human animals has relied on training the animal to give a behavioural cue when they have successfully discriminated (for a review, see [9]). Our work builds on that of Baru [21], who trained dogs to discriminate between synthesised vowels [a] and [i] through recognising formants as patterns, and responding by lifting a corresponding paw. However, as Baru's result used only synthesised voices and required the dogs to participate in up to 400 conditioning / reinforcement trials with negative reinforcement electric shocks to achieve accuracy, this level of discrimination was unlikely to represent a spontaneous ability in dogs [21]. Other experiments using natural voices have demonstrated that such diverse species as zebra finches (*Taeniopygia guttata*) [22] and chinchillas (*Chinchilla lanigera*) [23], among others, can be trained to normalise speaker differences to discriminate vowels. However, these studies too do not represent spontaneous responses as the research paradigms likewise relied on trained behaviours to indicate discrimination. Here, we measured spontaneous responses to natural voice stimuli in a habituation-dishabituation experiment and found that dogs did not require special nor extensive training to spontaneously normalise speakers and vowels.

Speech perception depends on the ability to parse relatively small differences in sounds and recognise these as meaningful [24]. Originally, it was believed that speech production and speech perception were inextricably linked abilities, and that perception required the brain to create a mental model of the articulatory gestures that produced the speech to recognise and categorise the sounds [24]. This "motor theory" posited that speech perception was unique to modern humans, as earlier hominins and other animals could not articulate their vocal apparatus to produce speech sounds and therefore could not make the mental connection between articulatory motions and the perceived sounds [25,26]. However, Kuhl and Miller [23] hypothesised that the two mechanisms of production

242 and perception are in fact separate, and, furthermore, suggested that speech perception
243 may at least be partly independent of speech production. This was based on evidence that
244 the ability to perceive speech sound differences is present in both very young human infants
245 (<1 month old) and also nonhuman animals including chinchillas, neither of which can
246 produce normal speech sounds [23,27]. Thus, their General Auditory Ability hypothesis
247 decoupled perception from production and suggested that humans have evolved speech
248 which can exploit existing perceptual categories rather than originating new abilities
249 [23,27]. Because dogs are not capable of speech production, our result that dogs can
250 normalise speaker differences to categorise vowels from formants lends some support to
251 this theory, suggesting that the ability to perform speaker normalisation may be a latent
252 ancestral trait. However, as dogs have undergone a long period of domestication of at least
253 13,000 years [28], it is possible that these normalisation abilities result from artificial
254 selection by humans for dogs that were more responsive to human vocal cues. Testing
255 speaker normalisation abilities in captive grey wolves (*Canis lupus*) that do not share the
256 same domestication history may help to clarify this point.
257 We also show that dogs can spontaneously discriminate between unfamiliar human
258 voices, even when the words spoken are not meaningful to the dogs, on the basis of very
259 limited exposure to just four words. This builds on previous results for familiar voice
260 recognition by both dogs [13] and cats [29,30]. Further investigations could establish which
261 aspects of the human voice are most important for the dogs' perception of speaker identity,
262 and what the effects of changing language, pitch, or other forms of speech modulation have
263 on dogs' perceptions of speaker identity. It is known that wolves can recognise familiar
264 conspecifics from their howls [17] and that dogs can recognise familiar humans by their

speech [13], but it has not yet been established if this cross-species ability was present in wolves or was specifically selected for during the domestication process.

In conclusion, dogs were found to spontaneously discriminate between both phonemic and identity cues in human speech. Dogs normalised differences in vocal production between same-gender speakers to recognise vowels and they could also use these differences to help to discriminate between unfamiliar speakers within genders. Thus, spontaneous speaker-normalisation to recognise vowels from formant patterns is not a uniquely human trait.

Acknowledgements

We thank Raystede Centre for Animal Welfare, RSPCA Mount Noddy Animal Centre, and all the dog owners for their assistance during testing. We also thank Harriet Grace, Imogen Fallon, Josephine McCartney, Sandra Bendoriute, Alice Keable, Jemma Forman, and Louise Brown for their assistance in data collection. We are grateful to Katarzyna Pisanski, Livio Favaro, and Matilde Massenet for commenting on earlier versions of this manuscript.

References

1. Titze IR. 1989 Physiologic and acoustic differences between male and female voices. *J. Acoust. Soc. Am.* (doi:10.1121/1.397959)

2. Fitch WT, Hauser MD. 2003 Unpacking " Honesty ": Vertebrate Vocal Production and Evolution of Acoustic Signals. *Acoust. Commun.* , 65–137. (doi:10.1007/0-387-22762-

8_3)

3. Kuhl PK. 1983 Perception of auditory equivalence classes for speech in early infancy. *Infant Behav. Dev.* **6**, 263–285. (doi:10.1016/S0163-6383(83)80036-8)

4. Fitch WT, Giedd J. 1999 Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *J. Acoust. Soc. Am.* (doi:10.1121/1.427148)

5. Childers DG, Wu K. 1991 Gender recognition from speech. Part II: Fine analysis. *J. Acoust. Soc. Am.* (doi:10.1121/1.401664)

6. Kreiman J, Sidtis D. 2011 Recognizing speaker identitiy from voice: Theoretical and ethological perspectives and a psychological model. *Found. Voice Stud. An Interdiscip. Approach to Voice Prod. Percept.*

7. Owren MJ, Cardillo GC. 2006 The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *J. Acoust. Soc. Am.* (doi:10.1121/1.2161431)

8. Liberman AM. 1982 On finding that speech is special. *Am. Psychol.* **37**, 107–144. (doi:10.1037//0003-066X.37.2.148)

9. Kriengwatana B, Escudero P, Cate C ten. 2015 Revisiting vocal perception in non-human animals: A review of vowel discrimination, speaker voice recognition, and speaker normalization. *Front. Psychol.* **6**. (doi:10.3389/fpsyg.2015.00543)

10. Fukuzawa M, Mills DS, Cooper JJ. 2005 More than just a word: Non-semantic

command variables affect obedience in the domestic dog (Canis familiaris). *Appl. Anim. Behav. Sci.* **91**, 129–141. (doi:10.1016/j.applanim.2004.08.025)

11. Fukuzawa M, Mills DS, Cooper JJ. 2005 The effect of human command phonetic characteristics on auditory cognition in dogs (Canis familiaris). *J. Comp. Psychol.* **119**, 117–120. (doi:10.1037/0735-7036.119.1.117)

12. Coutellier L. 2006 Are dogs able to recognize their handler's voice? A preliminary study. *Anthrozoos* **19**, 278–284. (doi:10.2752/089279306785415529)

13. Adachi I, Kuwahata H, Fujita K. 2007 Dogs recall their owner's face upon hearing the owner's voice. *Anim. Cogn.* **10**, 17–21. (doi:10.1007/s10071-006-0025-8)

14. Bachorowski J-A a, Owren MJ. 1999 Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *J. Acoust. Soc. Am.* **106**, 1054–1063. (doi:10.1121/1.427115)

15. Reby D, Hewison M, Izquierdo M, Pépin D. 2001 Red deer (Cervus elaphus) hinds discriminate between the roars of their current harem holder stag and those of neighbouring stags. *Ethology* **959**, 951–960.

16. Charlton BD, Ellis WAH, Larkin R, Tecumseh Fitch W. 2012 Perception of size-related formant information in male koalas (Phascolarctos cinereus). *Anim. Cogn.* **15**, 999–1006. (doi:10.1007/s10071-012-0527-5)

17. Font E, Carazo P, Márquez R, Palacios V, Font E, Marquez R, Carazo P. 2015 Recognition of familiarity on the basis of howls: a playback experiment in a captive

group of wolves. *Behaviour* **152**, 593–614. (doi:Doi 10.1163/1568539x-00003244)

18. Molnár C, Pongrácz P, Faragó T, Dóka A, Miklósi Á. 2009 Dogs discriminate between barks: The effect of context and identity of the caller. *Behav. Processes* **82**, 198–201. (doi:10.1016/j.beproc.2009.06.011)

19. Charlton BD, Reby D, McComb K. 2007 Female red deer prefer the roars of larger males. *Biol. Lett.* **3**, 382–385. (doi:10.1098/rsbl.2007.0244)

20. Peterson GEG., Barney HLH. HLH. 1952 Control Methods Used in a Study of the Vowels. *Jarnal Acoust. Soc. Am.* **24**, 175–184. (doi:10.1121/1.1906875)

21. Baru A V. 1975 Discrimination of synthesized vowels [a] and [i] with varying parameters (fundamental frequency, intensity, duration and number of formants) in dog. In *Auditory Analysis and Perception of Speech* (eds G Fant, MAA Tatham), pp. 91–101. London: ACADEMIC PRESS LTD.

22. Ohms VR, Gill A, van Heijningen CAA, Beckers GJL, ten Cate C. 2010 Zebra finches exhibit speaker-independent phonetic perception of human speech. *Proc Biol Sci* **277**, 1003–1009. (doi:10.1098/rspb.2009.1788)

23. Kuhl PK, Miller JD. 1975 Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants. *Science (80-.).* **190**, 69 LP – 72. (doi:10.1126/science.1166301)

24. Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. 1967 Perception of the speech code. *Psychol. Rev.* **74**, 431–461. (doi:10.1037/h0020279)

345 25. Mattingly IG. 1972 Speech cues and sign stimuli. *Am. Sci.*

346 26. Liberman AM, Mattingly IG. 1985 The motor theory of speech perception revised.
347 *Cognition* (doi:10.1016/0010-0277(85)90021-6)

348 27. Kuhl PK. 1988 Auditory perception and the evolution of speech. *Hum. Evol.* **3**, 19–43.
349 (doi:10.1007/BF02436589)

350 28. Thalmann O *et al.* 2013 Complete mitochondrial genomes of ancient canids suggest a
351 European origin of domestic dogs. *Science (80-.).* **342**, 871–874.
352 (doi:10.1126/science.1243650)

353 29. Saito A, Shinozuka K. 2013 Vocal recognition of owners by domestic cats (Felis catus).
354 *Anim. Cogn.* **16**, 685–690. (doi:10.1007/s10071-013-0620-4)

355 30. Takagi S, Arahori M, Chijiiwa H, Saito A, Kuroshima H, Fujita K. 2019 Cats match voice
356 and face: cross-modal representation of humans in cats (Felis catus). *Anim. Cogn.*
357 (doi:10.1007/s10071-019-01265-2)

358 31. Root-Gutteridge H, Ratcliffe VF, Korzeniowska A, Reby D. 2019 Data from: Dogs
359 perceive and spontaneously normalise formant-related speaker and vowel
360 differences in human speech sounds. (doi:https://doi.org/10.5061/dryad.g22g0dg)

361